



POLITECNICO DI BARI

DIPARTIMENTO DI INGEGNERIA ELETTRICA E DELL'INFORMAZIONE
CORSO DI LAUREA IN INGEGNERIA INFORMATICA E DELL'AUTOMAZIONE

TESI DI LAUREA
IN
METODI DI OTTIMIZZAZIONE

APPLICAZIONE DI TECNICHE DI INTELLIGENZA ARTIFICIALE
NELL'AMBITO DELLA GIUSTIZIA PENALE

Relatore:
Chiar.mo Prof. Ing. Tiziano POLITI

Laureando: Alessandro Cometa

ANNO ACCADEMICO 2023/2024



**LIBERATORIA ALLA CONSULTAZIONE DELLA TESI DI LAUREA DI CUI
ALL'ART.4 DEL REGOLAMENTO DI ATENEO PER LA CONSULTAZIONE DELLE
TESI DI LAUREA (D.R. n. 479 del 14/11/2016).**

Il sottoscritto Cometa Alessandro matricola 571060

Corso Ingegneria Informatica e dell'Automazione Laurea

autore della presente tesi di Laurea dal titolo
Applicazione di tecniche di intelligenza artificiale nell'ambito della giustizia penale.

Parola chiave intelligenza artificiale, metriche di performance, PCA, deviazione e densità

Abstract La tesi intende estrarre ulteriori informazioni nascoste nel dataset di partenza, come le componenti principali in cui è possibile scomporre il dataset stesso, e quale sia il miglior predittore applicabile. La tesi commenta anche l'etica deducibile dai risultati.

Autorizza

Non autorizza

la consultazione della presente tesi, fatto divieto a chiunque di riprodurre in tutto o in parte quanto in essa contenuto.

Bari,

Firma Cometa Alessandro

dedica

Ai parenti. Anche quelli che non ci sono più.

Indice

| | |
|---|-----|
| frontespizio | i |
| liberatoria | i |
| dedica | iii |
| Indice | v |
| Introduzione | vii |
| 1 Contesto: la giurisprudenza sul caso "Loomis vs Wisconsin" | 1 |
| 1.1 I soggetti, l'imputazione e la confutazione | 1 |
| Il giudizio di primo grado | 1 |
| Il giudizio di secondo grado | 1 |
| La sentenza della Corte Suprema | 2 |
| 1.2 Il ruolo del software COMPAS: la stima della probabile recidività | 2 |
| 1.3 Il libero convincimento del giudice in Italia | 3 |
| Sull'art. 115 c.p.c. | 3 |
| Sull'art. 116 c.p.c. | 4 |
| Sul giusto processo | 5 |
| Sulla figura dell' <i>amicus curiae</i> | 6 |
| 2 Metodologia | 7 |
| 2.1 Struttura di una AI | 7 |
| I dati in ingresso - L'input | 7 |
| I dati in uscita - L'output | 7 |
| I dati nel mentre - L'elaborazione | 7 |
| 2.2 L'importanza dei dati | 8 |
| Sulla normalizzazione delle tabelle | 8 |
| Sul preprocessing dei dati | 12 |
| 2.3 Possibili fonti dell'errore e dell'affidabilità di una AI | 13 |
| Sistema, modello, elaborazione | 13 |
| L'errore: patologia dell'apprendimento automatico ma non solo | 15 |
| Altri indici di affidabilità | 16 |
| 2.4 Funzionamento di una AI: i modelli | 17 |
| La rete neurale: il ruolo della funzione di attivazione | 17 |
| La regressione logistica: la rete neurale minima | 17 |
| K-nearest neighbors: sulle metriche di prossimità | 17 |

| | |
|---|-----------|
| Random forest: il modello esplicabile | 18 |
| Support Vector Machines: il metodo kernelizzato | 20 |
| 2.5 Le pipelines e le metriche di performance | 20 |
| 3 Conclusioni | 21 |
| 3.1 Presentazione dei risultati | 21 |
| 3.2 Un'ultima indagine sui dati: la PCA | 22 |
| 3.3 Sull'evoluzione della nozione di deviante | 24 |
| Bibliografia | 28 |
| Appendici | 29 |
| A test_sqlite.py | 30 |
| B test_sklearn.py | 33 |
| ringraziamenti | 36 |

Indice delle figure e delle tavole

| | | |
|-----|--|----|
| 2.1 | Modello classico: il sistema come "black-box" | 7 |
| 2.2 | Understanding the bias-variance tradeoff | 15 |
| 2.3 | L'albero migliore della Foresta, generata con i dati di COMPAS, poi potato a varie profondita (depth=4, depth=2) | 19 |
| 3.1 | PCA con i tre preprocessori | 23 |
| 2.1 | I principali operatori di join | 10 |
| 3.1 | Model Performance Comparison | 21 |

Introduzione

Il capitolo 1 illustra brevemente la genesi giuridico-penale-processuale dei dati di cui la tesi tratta, cioè il contesto da cui i dati provengono.

Il capitolo 2 esplora varie tecniche di algebra relazionale, di statistica inferenziale e di intelligenza artificiale con cui è possibile analizzare i dati, e le metriche di performance possibili sia per comparare e valutare i modelli, che per estrarre la più efficiente concatenazione di domande a risposta chiusa necessaria per comprendere meglio la struttura del dataset.

La tesi si conclude con un ulteriore capitolo dove vengono descritti i risultati ottenuti, ovvero si viene a scoprire quale sia il modello migliore e quali siano i dati più rilevanti nel dataset tramite tecniche di riduzione della dimensionalità del dataset di partenza, con ulteriori considerazioni in coda di natura sociologica, poiché il tema della predizione del reo, del reato e del recidivo non sembra essere un tema nuovo.

Capitolo 1

Contesto: la giurisprudenza sul caso "Loomis vs Wisconsin"

1.1 I soggetti, l'imputazione e la confutazione

Il giudizio di primo grado In sintesi, Eric L. Loomis nel 2013 viene fermato dalla polizia mentre era alla guida di un'automobile precedentemente usata per una sparatoria nello stato del Wisconsin, USA. Allo stato di fermo, al sig. Loomis vengono imputati vari capi d'accusa, tutti in recidiva, e con questi si va per i vari gradi di giudizio previsti dall'ordinamento giudiziario statunitense. All'attenzione del giudice, viene presentato anche il parere di un'intelligenza artificiale, il software COMPAS (Correctional Offender Management Profiling for Alternative Sanctions), perché la produzione di un PSI (Presentence Investigation Report) era stata richiesta dalla Corte stessa, dopo che questa aveva recepito dal sig. Loomis l'ammissione di colpevolezza.¹

Il giudizio di secondo grado L'uso di COMPAS nella determinazione della pena - nel giudizio di primo grado - violava il diritto all'equo processo sotto tre profili, secondo la difesa del sig. Loomis:

1. il diritto ad essere condannato ad una determinata pena sulla base di informazioni accurate, delle quali non si poteva disporre in quanto coperte da diritti di proprietà industriale
2. il diritto di essere condannato ad una pena individualizzata
3. l'uso improprio del dato di genere nella determinazione della pena.

Nessuno dei tre punti viene accettato dall'analogo USA della nostra Corte d'Appello:

1. pur vigendo il segreto industriale sul funzionamento di COMPAS, al sig. Loomis è stata data la possibilità di verificare i dati forniti in input, e di contestare il risultato finale del calcolo - cosa che il sig. Loomis non ha fatto;

¹Un racconto più approfondito dei fatti del caso è presente in Giurisprudenza Penale, ad opera di Stefania Carrer[1].

2. sul diritto del reo ad una pena individualizzata, questo non è stato violato da COMPAS, perché se anche COMPAS non dovesse tener conto delle peculiarità dell'individuo - cosa di cui non si può sapere, poiché il funzionamento interno è secretato - al giudice è comunque dato di scegliere in totale discrezionalità dati puri o aggregati provenienti anche da fonti diverse;
3. per quanto il software effettivamente dia maggior peso a certi fattori contro altri (come verrà dimostrato anche in questa tesi), la Corte ribadisce che l'accuratezza della valutazione dipende da un monitoraggio costante, aggiornato sulla base dei cambiamenti sociali (come si vedrà in seguito, anche varie nozioni sociologiche, come quelle di devianza e di deviante, si sono evolute nel tempo, evoluzione tuttora in corso) - all'utente di COMPAS è prescritta una cautela speciale.

La sentenza della Corte Suprema La Corte Suprema (che è l'analogo USA della nostra Corte di Cassazione) arriva nel 2016 a sostenere, in una discussa sentenza, quali limiti e cautele devono essere applicati quando un giudice ha a che fare con pareri provenienti da software simili a COMPAS, nonché da COMPAS stesso.

Si è infatti stabilito che tali software possono essere considerati fattori rilevanti in questioni quali

1. la comminazione di misure alternative alla detenzione per gli individui a basso rischio di recidiva;
2. la valutazione della possibilità di controllare un criminale in modo sicuro all'interno della società, anche con l'affidamento in prova;
3. l'imposizione di termini e condizioni per la libertà vigilata, la supervisione e per le eventuali sanzioni alle violazioni delle regole previste dai regimi alternativi alla detenzione.

Ribadendo la necessità che il giudice applichi i risultati COMPAS facendo esercizio della propria discrezionalità sulla base del bilanciamento con altri fattori, la Corte ha confermato che l'uso dello strumento non può riguardare il grado di severità della pena sulla base di circostanze attenuanti od aggravanti, né la decisione sull'incarcerazione dell'imputato. Ha specificato infatti che lo scopo di COMPAS è quello di individuare le esigenze del soggetto che deve scontare la pena e di valutare il rischio di reiterazione del reato.^[1]

1.2 Il ruolo del software COMPAS: la stima della probabile recidività

La sentenza ha suscitato vasto interesse accademico e giornalistico sulla possibilità di un software in generale e di un'intelligenza artificiale in particolare, notoriamente non umani, di interferire nella decisione dosimetrica, tutta umana, del giudice che sta per pronunciarsi su un caso penale. ^[2] ^[3] ^[4] ^[5] ^[6]

Difficile non citare Ugo Pagallo, Luciano Floridi e Giovanni Sartori come persone che in Italia si sono spese sul tema più ampio dell'interazione tra uomo e intelligenza artificiale, sia sul particolare piano giuridico, sia in vari altri campi del vivere per come lo conosciamo oggi.

La sentenza arriva anche all'attenzione di autori di un articolo su Science[7], dove si punta a lizza...

1.3 Il libero convincimento del giudice in Italia

Negli USA, nel frattempo, ProPublica, una testata giornalistica indipendente, scrive una serie di articoli mirati a criticare l'efficacia statistica del software COMPAS.[8]

La sentenza della Corte Suprema ha avuto origine, tra le altre cose, dalla supposizione che il parere di COMPAS non fosse stato generato secondo la Costituzione effettivamente statuita². La Corte Suprema, dunque, procede stabilendo che questo parere è legittimo, anche se è utilizzabile solo in alcuni casi e per alcuni scopi. Potrebbe essere interessante esplorare, a questo punto, se nell'ordinamento italiano potrebbe accadere qualcosa del genere: in Italia, una profilazione preliminare come quella offerta dal software COMPAS verrebbe presa in considerazione dal giudice italiano di primo grado? La normativa di riferimento si ritrova al libro primo (dedicato alle disposizioni generali), titolo V ("Dei poteri del giudice") del codice di procedura civile, agli artt. 115 e 116.

Sull'art. 115 c.p.c.

Salvi i casi previsti dalla legge³, il giudice deve porre a fondamento della decisione le prove proposte dalle parti o dal pubblico ministero nonché i fatti non specificatamente contestati dalla parte costituita⁴.

Il giudice può tuttavia, senza bisogno di prova, porre a fondamento della decisione le nozioni di fatto che rientrano nella comune esperienza⁵.

Sentenze di Cassazione rilevanti sono:

- Cassazione civile, Sez. VI, ordinanza n. 3126 del 1 febbraio 2019

L'onere di contestazione riguarda le allegazioni delle parti e non le prove assunte, la cui valutazione opera in un momento successivo alla definizione dei fatti controversi ed è rimessa all'apprezzamento del giudice.

²Citando la formula di Hans Kelsen.

³In virtù del principio dispositivo sono le parti a proporre al giudice gli elementi di prova su cui basare il proprio convincimento. E' bene precisare che la legge prevede ipotesi eccezionali, in cui il giudice dispone ex officio mezzi di prova, come nel caso degli artt. 117 (interrogatorio non formale delle parti), 118 (ispezione di persone e di cose), 213 (richiesta di informazioni alla P.A.), 257 (assunzione di nuovi testimoni), 421, 442 (poteri istruttori del giudice in controversie di lavoro e di previdenza e di assistenza obbligatorie), 714 (poteri istruttori nei procedimenti di interdizione o inabilitazione).

⁴L'articolo in esame è stato aggiornato con le modifiche introdotte dalla Legge 18 giugno 2009, n. 69. Con tale nuova formulazione viene conferito al giudice il potere di ritenere provati, accanto ai fatti notori, anche quelli che non sono stati specificamente contestati dalla controparte né direttamente né indirettamente. La ratio di tale riforma è dettata dall'esigenza di attuare il disposto normativo di cui agli artt. 2697, 2698 c.c. che impongono a chi voglia far valere un fatto in giudizio di provarne i fatti che ne sono a fondamento: è il principio dell'onere della prova.

⁵L'ultimo comma rappresenta una deroga al principio dispositivo ed al contraddittorio in quanto introduce nel processo civile prove non fornite dalle parti e relative a fatti dalle stesse non vagliati né controllati. Tuttavia, la possibilità per il giudice di ricorrere ai c.d. fatti notori sussiste solo ed esclusivamente nel caso in cui si tratti di fatti acquisiti alle conoscenze della collettività, in un dato tempo e luogo, con tale grado di certezza da apparire indubitabile ed incontestabile. Tipico esempio di fatto notorio è la svalutazione monetaria.

CAPITOLO 1. CONTESTO: LA GIURISPRUDENZA SUL CASO "LOOMIS VS WISCONSIN"⁴

Qui già si apre a quanto disciplinato dall'articolo seguente: la valutazione delle prove e il libero apprezzamento delle stesse prove da parte del giudice.

Sull'art. 116 c.p.c.

Il giudice deve valutare le prove secondo il suo prudente apprezzamento⁶, salvo che la legge disponga altrimenti⁷.

Il giudice può desumere argomenti di prova dalle risposte che le parti gli danno a norma dell'articolo seguente, dal loro rifiuto ingiustificato a consentire le ispezioni che egli ha ordinate e, in generale, dal contegno delle parti stesse nel processo⁸.

Torna molto pratico menzionare anche alcune sentenze della Corte di Cassazione:

- Cassazione civile, Sez. I, sentenza n. 17392 del 1 settembre 2015

Nell'ordinamento processuale vigente manca una norma di chiusura sulla tassatività tipologica dei mezzi di prova, sicché il giudice può legittimamente porre a base del proprio convincimento anche prove cd. atipiche, quali le dichiarazioni scritte provenienti da terzi, della cui utilizzazione fornisca adeguata motivazione e che siano idonee ad offrire elementi di giudizio sufficienti, non smentiti dal raffronto critico con le altre risultanze istruttorie, senza che ne derivi la violazione del principio di cui all'art. 101 c.p.c., atteso che, sebbene raccolte al di fuori del processo, il contraddittorio si instaura con la produzione in giudizio.

- Cassazione civile, Sez. III, sentenza n. 13229 del 26 giugno 2015

Nel vigente ordinamento processuale, improntato al principio del libero convincimento del giudice e in assenza di una norma di chiusura sulla tassatività tipologica dei mezzi di prova, questi può porre a fondamento della decisione anche prove atipiche, non espressamente previste dal codice di rito, della cui utilizzazione fornisca adeguata motivazione e che siano idonee ad offrire elementi di giudizio sufficienti, non smentiti dal raffronto critico con le altre risultanze del processo.

- Cassazione civile, Sez. Lavoro, sentenza n. 13054 del 10 giugno 2014

⁶Per "prudente apprezzamento" si intende il compito del giudice tenuto a valutare la attendibilità di ogni circostanza posta alla sua attenzione, ma non necessariamente ad utilizzarla e che può poi anche considerare tutti gli elementi con efficacia probatoria emersi nel corso del giudizio.

⁷Il riferimento è quello alle c.d. prove legali, quali le prove documentali (atto pubblico e scrittura privata autenticata o riconosciuta) o quelle assunte nel processo come la confessione (v. 228 e ss.), il giuramento (v. 233 e ss.) e la testimonianza (v. 244 e ss.). Si tratta di prove la cui efficacia è predeterminata dalla legge e di fronte alle quali al giudice è impedita ogni valutazione sul contenuto della stessa, dovendosi semplicemente attenere alle risultanze della prova offerta, così come legalmente stabilito.

⁸La norma si può anche considerare riferita all'ipotesi in cui il giudice possa valutare, per la formazione del suo convincimento, anche prove formate in un diverso processo. Tali prove possono essere utilizzate dal giudice come semplici indizi idonei a fornire utili e concorrenti elementi di giudizio oppure possono avere valore di prova esclusiva come ad es. accade nel caso di una perizia svolta in sede penale o di una consulenza tecnica svolta in altra sede civile.

In tema di procedimento civile, sono riservate al giudice del merito l'interpretazione e la valutazione del materiale probatorio, nonché la scelta delle prove ritenute idonee alla formazione del proprio convincimento, con la conseguenza che è insindacabile, in sede di legittimità, il "peso probatorio" di alcune testimonianze rispetto ad altre, in base al quale il giudice di secondo grado sia pervenuto ad un giudizio logicamente motivato, diverso da quello formulato dal primo giudice.

- Cassazione civile, Sez. VI, sentenza n. 26550 del 12 dicembre 2011

Il giudice del merito può porre a fondamento della propria decisione una perizia stragiudiziale, anche se contestata dalla controparte, purché fornisca adeguata motivazione di tale sua valutazione, attesa l'esistenza, nel vigente ordinamento, del principio del libero convincimento del giudice.

Quindi, la possibilità che un report proveniente da un software come COMPAS possa essere utilizzato nel processo italiano è rimessa alla sensibilità del giudice sul tema.

Sul giusto processo Si applicano, in ogni caso, le norme sul giusto processo, ex art. 111 della Costituzione.

La giurisdizione si attua mediante il giusto processo regolato dalla legge.

Ogni processo si svolge nel contraddittorio tra le parti, in condizioni di parità, davanti a giudice terzo e imparziale. La legge ne assicura la ragionevole durata.

Nel processo penale, la legge assicura che la persona accusata di un reato sia, nel più breve tempo possibile, informata riservatamente della natura e dei motivi dell'accusa elevata a suo carico; disponga del tempo e delle condizioni necessari per preparare la sua difesa; abbia la facoltà, davanti al giudice, di interrogare o di far interrogare le persone che rendono dichiarazioni a suo carico, di ottenere la convocazione e l'interrogatorio di persone a sua difesa nelle stesse condizioni dell'accusa e l'acquisizione di ogni altro mezzo di prova a suo favore; sia assistita da un interprete se non comprende o non parla la lingua impiegata nel processo.

Il processo penale è regolato dal principio del contraddittorio nella formazione della prova. La colpevolezza dell'imputato non può essere provata sulla base di dichiarazioni rese da chi, per libera scelta, si è sempre volontariamente sottratto all'interrogatorio da parte dell'imputato o del suo difensore.

La legge regola i casi in cui la formazione della prova non ha luogo in contraddittorio per consenso dell'imputato o per accertata impossibilità di natura oggettiva o per effetto di provata condotta illecita.

Tutti i provvedimenti giurisdizionali devono essere motivati.

Contro le sentenze e contro i provvedimenti sulla libertà personale, pronunciati dagli organi giurisdizionali ordinari o speciali, è sempre ammesso ricorso in Cassazione per violazione di legge. Si può derogare a tale norma soltanto per le sentenze dei tribunali militari in tempo di guerra.

Contro le decisioni del Consiglio di Stato e della Corte dei conti il ricorso in Cassazione è ammesso per i soli motivi inerenti alla giurisdizione.

Sulla figura dell'amicus curiae A p. 42, Simone Paduanelli riferisce nella sua tesi di laurea magistrale all'Università di Milano-Bicocca [9] che si stanno tentando vari approcci alla tecnologia: sia tramite lavori preparatori di creazione di banche dati pubbliche ministeriali sulle sentenze civili; sia tramite partenariati interuniversitari che, per esempio, radunano le università locali lombarde sotto il progetto "Predictive Justice" (che dovrebbe terminare nel 2026), voluto dal presidente della Corte d'Appello di Brescia, il quale auspica una maggior diffusione della consapevolezza circa il rispetto degli specifici diritti rilevati dalle sentenze pregresse (il "case-law"). Anche l'università Sant'Anna di Pisa si sta attivando, in collaborazione con le corti di Genova e Pisa, per la creazione di una AI che si occupi di apporre annotazioni semantiche alle decisioni giudiziali, per favorire un recupero più rapido delle informazioni: il progetto attualmente sta restituendo dati soddisfacenti negli ambiti del diritto di famiglia e della responsabilità da danno personale.

Nulla di poi così simile ad un software di profilazione come COMPAS, ma dei tentativi di indagare le interazioni tra giurisprudenza e intelligenza artificiale sono in atto. Ciononostante, alcune considerazioni possono essere fatte sull'ingegneria dietro un software come COMPAS - se serve, immaginandolo come strumento a disposizione del diritto vigente in Italia⁹.

⁶Questo diritto, come da art. 10 Cost., estende il diritto vigente al diritto comunitario e al diritto internazionale.

Capitolo 2

Metodologia

2.1 Struttura di una AI

La canonica concettualizzazione a "black-box"¹, poiché generalissima, si applica a qualsiasi sistema, e quindi anche a qualsiasi intelligenza artificiale; queste proprietà apparterranno anche a COMPAS.

I dati in ingresso - L'input ProPublica è riuscita ad estrarre dei dati su alcuni casi su cui è stato prodotto il report PSI di cui si è parlato prima [10]. Sono state raccolte le 137 domande sulla base delle quali il sistema COMPAS era chiamato a produrre una stima numerica della probabilità che il soggetto si rendesse nuovamente reo.[11]

I dati in uscita - L'output L'output corrisponde ad un punteggio da 0.0 a 10.0, corrispondente alla probabilità stimata di recidività. Punteggi più alti corrispondono ad un reo a più probabile recidività.

I dati nel mentre - L'elaborazione Come le 137 domande portano ad un punteggio? Il materiale di partenza non permette di rispondere a questa domanda precisamente, tuttavia il file "compas.db" contiene la tabella "people", che sarà oggetto di indagine di questa tesi. La mancanza di competenze giuridiche adeguate sul processo penale statunitense obbliga a proiettare i dati su una restrizione di colonne.

¹Il disturbo è anche chiamato in letteratura "ingresso non controllabile".

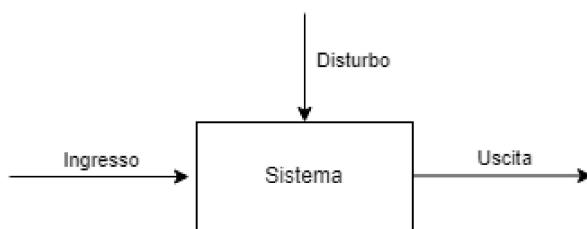


Figura 2.1: Modello classico: il sistema come "black-box"

2.2 L'importanza dei dati

Sulla normalizzazione delle tabelle Il file che ospita l'input, "compas.db", possiede numerose tabelle. L'attenzione gravita naturalmente, trattandosi di un'elaborazione statistica, sulla tabella "people", la quale è, abbastanza evidentemente, non normalizzata, se non nella cosiddetta "Prima forma normale".

La normalizzazione algebrica dei dati è una sequenza di accorgimenti volti al minimizzare i dati ridondanti, che evitano sia la perdita di dati rilevanti, sia eventuali confusioni nell'accesso ai dati.

Ispezione preliminare Il file che era stato originariamente preso in considerazione, `compas-scores.csv`, presentava colonne duplicate.

```
id,
name,
first,
last,
compas_screening_date,
sex,
dob,
age,
age_cat,
race,
juv_fel_count,
decile_score,
juv_misd_count,
juv_other_count,
priors_count,
days_b_screening_arrest,
c_jail_in,
c_jail_out,
c_case_number,
c_offense_date,
c_arrest_date,
c_days_from_compas,
c_charge_degree,
c_charge_desc,
is_recid,
num_r_cases,
r_case_number,
r_charge_degree,
r_days_from_arrest,
r_offense_date,
r_charge_desc,
r_jail_in,
r_jail_out,
is_violent_recid,
```

```
num_vr_cases,
vr_case_number,
vr_charge_degree,
vr_offense_date,
vr_charge_desc,
v_type_of_assessment,
v_decile_score,
v_score_text,
v_screening_date,
type_of_assessment,
decile_score,
score_text,
screening_date
```

Già ad una prima ispezione, ci si accorge che alcune colonne sono duplicate, infatti la colonna `decile_score` figura due volte. Con un'ispezione dei dati contenuti nel dataset, anche altre colonne risultano contenere, in ordine, gli stessi dati: hanno soltanto nomi diversi, e sono state eliminate in quanto il seguente blocco di codice non ha sollevato alcun `AssertionError`:

```
def eq_col(data, feat1, feat2):
    for x in range(0, 11757):
        assert data[feat1].iloc[x] == data[feat2].iloc[x]
    print(f'{feat1} coincide con {feat2}')
eq_col(data, "decile_score", "decile_score.1")
eq_col(data, "compas_screening_date", "v_screening_date")
eq_col(data, "compas_screening_date", "screening_date")
```

Sono, evidentemente, figlie di un qualche join infelice. Esiste una nozione che permetta di contrastare la produzione di non-valori che l'operatore di (outer-)join tende a mettere in atto? Ma, prima di tutto, cos'è il join?

Sull'algebra relazionale - L'operatore di join L'operatore di join prevede due insiemi di partenza che si intersecano tra loro. I loro nomi potrebbero essere A e B, senza perdita di generalità.

$$A \bowtie_{\langle cond \rangle} B = \sigma_{\langle cond \rangle} A \times B \quad (2.1)$$

| id_A | $Dato_1$ | $Dato_2$ | $Dato_3$ | id_B | $Dato_4$ | $Dato_5$ | $Dato_6$ |
|--------|----------|----------|----------|--------|----------|----------|----------|
| 1 | 6 | 87837 | 787 | 1 | 87 | 877 | 788 |
| 2 | 7 | 78 | 5415 | 4 | 978 | 995 | 6508 |
| 3 | 545 | 778 | 7507 | 5 | 6756 | 586 | 9507 |

$$A \bowtie_{\langle id_A=id_B \rangle} B = \{id, Dato_1, Dato_2, Dato_3, Dato_4, Dato_5, Dato_6\} \quad (2.2)$$

Su questi due insiemi possono essere effettuati vari tipi di join.² I più famosi sono:

| Tipo | id | $Dato_1$ | $Dato_2$ | $Dato_3$ | $Dato_4$ | $Dato_5$ | $Dato_6$ |
|------------------|------|----------|----------|----------|----------|----------|----------|
| Inner join | 1 | 6 | 87837 | 787 | 87 | 877 | 788 |
| Left outer join | 1 | 6 | 87837 | 787 | 87 | 877 | 788 |
| Right outer join | 1 | 6 | 87837 | 787 | 87 | 877 | 788 |
| Full outer join | 1 | 6 | 87837 | 787 | 87 | 877 | 788 |
| Inner join | 2 | 7 | 78 | 5415 | NaN | NaN | NaN |
| Left outer join | 2 | 7 | 78 | 5415 | NaN | NaN | NaN |
| Right outer join | 2 | 7 | 78 | 5415 | NaN | NaN | NaN |
| Full outer join | 2 | 7 | 78 | 5415 | NaN | NaN | NaN |
| Inner join | 3 | 545 | 778 | 7507 | NaN | NaN | NaN |
| Left outer join | 3 | 545 | 778 | 7507 | NaN | NaN | NaN |
| Right outer join | 3 | 545 | 778 | 7507 | NaN | NaN | NaN |
| Full outer join | 3 | 545 | 778 | 7507 | NaN | NaN | NaN |
| Inner join | 4 | NaN | NaN | NaN | 978 | 995 | 6508 |
| Left outer join | 4 | NaN | NaN | NaN | 978 | 995 | 6508 |
| Right outer join | 4 | NaN | NaN | NaN | 978 | 995 | 6508 |
| Full outer join | 4 | NaN | NaN | NaN | 978 | 995 | 6508 |
| Inner join | 5 | NaN | NaN | NaN | 6756 | 586 | 9507 |
| Left outer join | 5 | NaN | NaN | NaN | 6756 | 586 | 9507 |
| Right outer join | 5 | NaN | NaN | NaN | 6756 | 586 | 9507 |
| Full outer join | 5 | NaN | NaN | NaN | 6756 | 586 | 9507 |

Tabella 2.1: *I principali operatori di join*

Il file `compas-scores.csv` sembra essere stato prodotto da una left outer join tra varie tabelle, poiché se di tutte le persone si danno le informazioni anagrafiche, solo una minoranza è stata autrice di reati violenti: per tutti gli altri, alle celle corrispondenti, si dà NaN - Not a Number. Come tentare una ricostruzione delle tabelle originali?

²Per una trattazione più estesa, si veda il par. 3.1.5 di [12]

Sull'algebra relazionale - Le forme normali Qualche definizione informale:

1. Prima forma normale (1NF):

Si richiede che nessuna riga sia uguale alle altre, e che nessuna colonna contenga dati compositi.

2. Seconda forma normale (2NF):

In aggiunta a quanto previsto dalla 1NF, si richiede che ciascun attributo non-chiave dipenda sull'intera chiave - ovvero che ogni attributo non-chiave abbia una piena dipendenza funzionale su ciascuna chiave candidata.

Esempio:

Inventario = {Prodotto, Magazzino, Prezzo, NomeProdotto, IndirizzoMagazzino};

La tabella non è in 2NF perché, anche ammesso che sia in 1NF, alcuni attributi non sono funzionalmente dipendenti dall'interezza della chiave primaria.

La decomposizione della tabella che produce tabelle in 2NF porterà alla seguente situazione:

Inventario = {Prodotto, Magazzino, Prezzo}

Prodotti = {Prodotto, NomeProdotto}

Magazzini = {Magazzino, IndirizzoMagazzino}

3. Terza forma normale (3NF):

Oltre a quanto previsto dalla 2NF, non devono essere presenti dipendenze funzionali transitive tra gli attributi della tabella, ovvero gli attributi devono tutti dipendere dalla chiave primaria della tabella.

Per "chiave" si intende, come nell'informatica delle basi di dati, quell'insieme di informazioni che identificano completamente l'elemento presente nella base di dati. Tramite questo recupero dell'oggetto, quindi, è possibile conoscerne anche le altre informazioni.³ Per quanto riguarda la tabella esposta dal file `compas-scores.csv`, la tabella non è nemmeno in 1NF, perché i dati contenuti nella colonna "name" non sono atomici.⁴

³Per una trattazione più estesa sulle forme normali, si veda il cap. 9 di [12]

⁴La natura dei dati non ha permesso un'agevole decomposizione, pertanto si è proceduto effettuando una proiezione dello spazio dei dati su uno spazio a minore dimensionalità, selezionando solo alcune colonne ritenute più rilevanti tra quelle nella tabella "people" del file `compas.db`, tabella da cui sembra essere stato generato il file `compas-scores.csv`.

Sul preprocessing dei dati

Gestione dei dati non numerici Alcuni dati non sono in forma numerica, mentre ciascuno dei modelli prende in ingresso solo dati numerici, pertanto è stato necessario individuare strategie per gestirli. Il primo problema è stato sui nomi dei condannati: risolto esportando quei dati in una tabella a parte, con l'id a fungere sia da chiave primaria che da chiave esterna. Dopodiché, le date: tutte convertite nella differenza tra la data in questione e una data di riferimento. Per la data di riferimento, tre opzioni:

- La scelta convenzionale, il 01/01/1970, ovvero il cosiddetto “Unix time”.
- La scelta più logica, la data più antica presente nel dataset: 14/10/1919, che garantisce l'assenza di valori negativi.
- La scelta effettivamente selezionata: anche per mostrare che una data vale l'altra, la data selezionata è il 24/6/1936. È sufficientemente buona, solo 5 date nell'intero dataset sono anteriori a questa.

Infine, un altro problema è quello delle colonne che ospitano dati categorici. Si è proceduto con una tecnica, chiamata One-hot-encoding, che genera una colonna per ciascuna categoria, assegnando 1 nella nuova cella appartenente alla riga dove il dato originario corrisponde alla colonna della cella, e 0 alle celle della riga in tutte le altre nuove colonne. La colonna “race” è stata trattata così. Ad una più attenta analisi, la colonna “sex” contiene solo dati “Male” e “Female” (è stato verificato), pertanto è stata trattenuta solo la colonna “is_male”: la colonna “is_female” sarebbe stata ridondante, in quanto calcolabile all'occorrenza negando il dato di fatto booleano della colonna “is_male”.

Ora, ciascuna di queste colonne presenta dei massimi e dei minimi che variano da una colonna all'altra. Questo non è positivo per i modelli, perché $\text{span} (= x_{\max} - x_{\min})$ più elevati faranno credere al modello che un dato sia più importante di altri, mentre magari è solo un risultato indesiderato delle codifiche di cui sopra. Pertanto, tre tecniche di riscalatura:

- Standard scaling: si calcola, per ogni dato in input, un punteggio

$$z = \frac{x - \mu}{\sigma} \quad (2.3)$$

Questa tecnica ha il compito di rendere unitaria la varianza, e nulla la media, della collezione di dati che gli si dà in input.

- Minmax scaling

$$z = \frac{x - \min}{\max - \min} \quad (2.4)$$

Questa tecnica garantisce che tutti i dati siano mappati nell'intervallo $[0, 1]$.

- Quantile transformation: una funzione di distribuzione cumulativa mappa i valori originali ad una distribuzione uniforme. Questa tecnica, poiché distribuisce in maniera uniforme i dati, è particolarmente efficiente nella reiezione dei disturbi provocati da dati isolati (outliers).

2.3 Possibili fonti dell'errore e dell'affidabilità di una AI

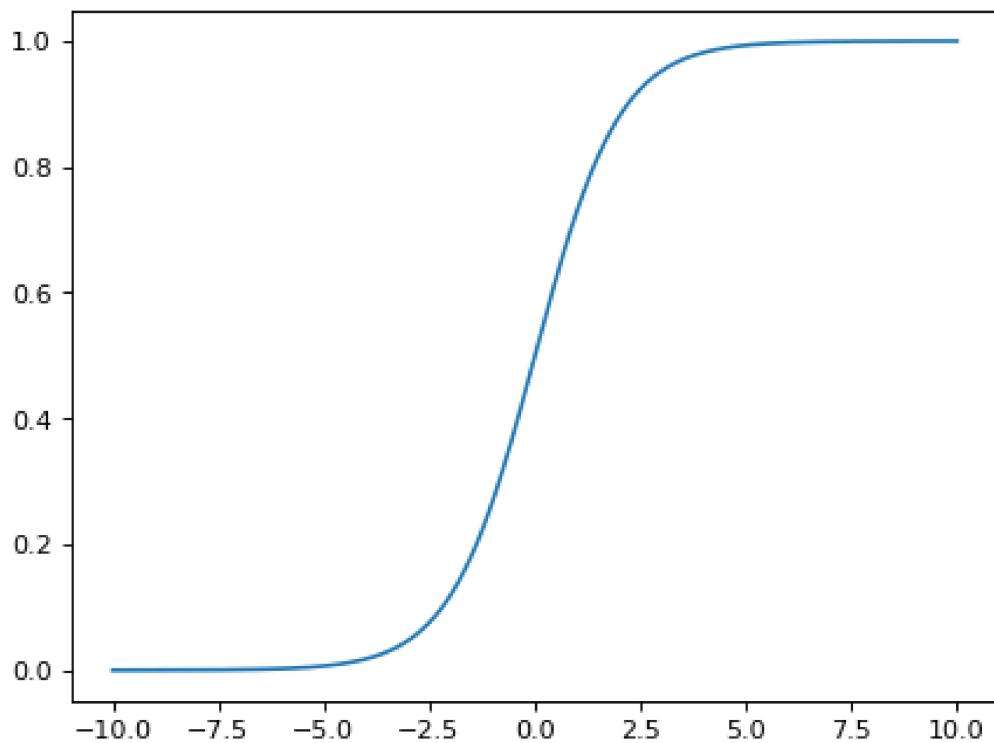
Sistema, modello, elaborazione

La pipeline La libreria selezionata è scikit-learn (versione 1.5.2). Scikit-learn espone vari strumenti per l'analisi dei dati, tra cui:

1. preprocessing.StandardScaler
2. preprocessing.MinMaxScaler
3. preprocessing.QuantileTransformer
4. preprocessing.scale()
5. preprocessing.minmax_scale()
6. preprocessing.quantile_transform()
7. linear_model.LogisticRegression
8. neighbors.KNearestNeighbors
9. ensemble.RandomForestRegressor
10. svm.SVR
11. model_selection.train_test_split()
12. pipeline.make_pipeline()
13. decomposition.PCA
14. metrics.r2_score()
15. metrics.mean_squared_error()
16. metrics.make_scorer()
17. tree.plot_tree()

Le prime 6 risorse sono state spiegate nel paragrafo precedente, dedicato al preprocessing dei dati. Ora ci si concentra sul resto, ovvero sui modelli resi disponibili dalla libreria. La regressione logistica è un processo di apprendimento automatico per cui i dati in input vengono scrutinati da una funzione detta logistica, o sigmoide:

$$f(x) = \frac{1}{1 + e^{-x}} \quad (2.5)$$



Come da foto, la funzione è monotona crescente, e ha valori nell'intervallo $[0, 1]$. In particolare, i valori estremali 0 e 1 sono asintoti per la funzione. Output ragionevolmente vicini a questi due valori, in base a delle soglie di tolleranza predeterminate, permettono di effettuare una classificazione o una regressione. La funzione è interessante per il fatto di essere derivabile.

$$f'(x) = f(x)(1 - f(x)) \quad (2.6)$$

L'errore: patologia dell'apprendimento automatico ma non solo

Bias-variance tradeoff L'errore quadratico medio è scomponibile in tre contributi additivi:⁵

- La varianza (o errore casuale): $E_{X,D}[(h_D(X) - \bar{h}_D(X))^2]$;
- Il quadrato del bias (o errore sistematico): $E_X[(\bar{h}(X) - \bar{y}(X))^2]$;
- Il rumore (ineliminabile) dei dati: $E_{X,y}[(\bar{y}(X) - y)^2]$;

La qualità di un modello è valutabile dalla sua capacità di minimizzare l'errore complessivo. Un primo problema risiede nel fatto che, man mano che cresce la complessità del modello, benché cali l'errore sistematico, cresce l'errore casuale. Il modello, quindi, benché interessato programmaticamente a minimizzare l'errore, non potrà mai annullarlo.

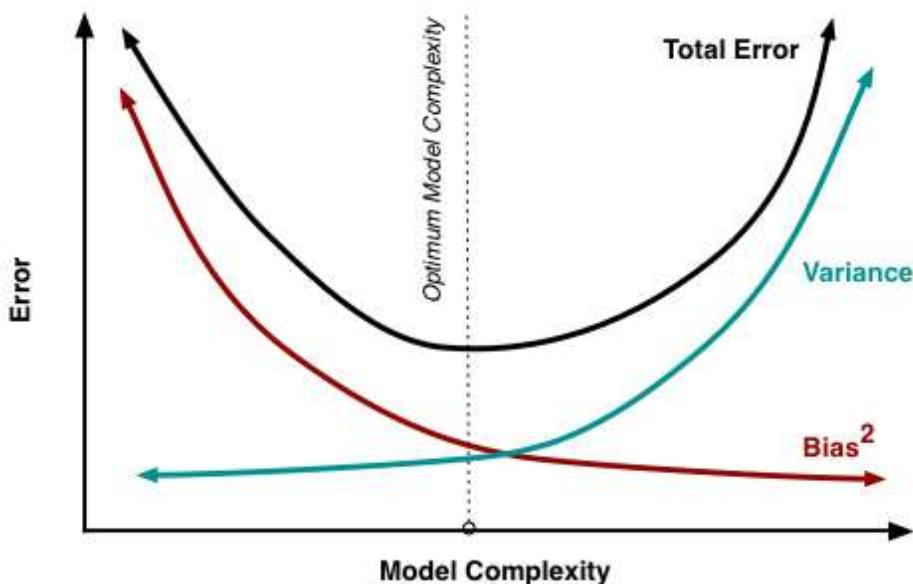


Figura 2.2: Understanding the bias-variance tradeoff

Overfitting e underfitting

- Il fenomeno dell'overfitting si manifesta quando il sistema intelligente non riesce a generalizzare, e quindi ad apprendere alcunché dai dati. Ciò implica che crea un modello troppo più complesso e rigido del necessario sui dati presenti nel training set, talmente rigido da non comportarsi correttamente coi dati nel test set.
- Il fenomeno dell'underfitting, dualmente, coincide con modelli semplicistici, altrettanto inadeguati, con la conseguenza che il modello espone errore anche nei dati del training set, diminuendo la varianza in favore dell'errore sistematico (e del suo quadrato).

⁵Per approfondire, si veda [13]

Altri indici di affidabilità

Matrici di confusione La matrice di confusione, detto di un sistema di classificazione binaria (esempi: vero/falso, alto/basso, spam/non-spam) è una matrice con due righe (positivo attuale e negativo attuale) e due colonne (positivo predetto e negativo predetto). Ciascuna delle celle contiene il conteggio dei risultati:

- davvero positivi (TP);
- davvero negativi (TN);
- erroneamente rilevati come positivi (FP);
- erroneamente rilevati come negativi (FN).

Precision e recall La bontà di un modello emerge, ad un rapido calcolo, da alcuni indici di semplice computazione, tra cui "precision" e "recall".

$$\text{Precision} = TP / (TP + FP) \quad (2.7)$$

$$\text{Recall} = TPR = TP / (TP + FN) \quad (2.8)$$

Più sono alti i valori di questi indici, migliore è il modello:

- "precision" si occupa di verificare quante, tra le predizioni positive, sono corrette;
- "recall" tiene traccia di quanti, tra i veri positivi, sono stati predetti correttamente.

Curva ROC Una rappresentazione più interessante, specie nell'ottica di comparare più predittori, richiede di considerare un ulteriore indice:

$$FPR = FP / (TN + FP) \quad (2.9)$$

La curva ROC (Receiver Operating Characteristic, o anche Relative Operating Characteristic) esiste in uno spazio cartesiano dove su un asse è rappresentato il TPR, e sull'altro il FPR.

Un predittore binario è tanto più efficace, quanto più la sua caratteristica operativa si distanzia da quella di un decisore casuale, che avrebbe invece un tasso di veri positivi identica al tasso di falsi positivi.

2.4 Funzionamento di una AI: i modelli

La rete neurale: il ruolo della funzione di attivazione La funzione di attivazione $\phi(x)$ è, data una rete neurale, quella funzione che trasforma tutti gli input di un percettrone in un dato numero di output. Il percettrone è ispirato al neurone, componente elementare del cervello umano, che riceve più segnali elettrochimici dai collegamenti sinaptici con altri neuroni. Si è detto prima che, per un funzionamento ottimale del modello, l'errore è necessario, e per evitare un errore casuale troppo alto (che pregiudicherebbe la ripetibilità dei risultati), viene preso in considerazione un errore sistematico fisso adeguato, il "bias".

$$f(x) = b + w_1x_1 + w_2x_2 + \dots + w_nx_n \quad (2.10)$$

$$g(x) = \phi(f(x)) \quad (2.11)$$

Funzioni di attivazione alternative: descrizione analitica Tra le funzioni di attivazione $\phi(x)$ che si possono utilizzare si segnalano:

- **Sigmoide** (che è stata già spiegata)

- **ReLU**

La Rectified Linear Unit ha una espressione molto semplice:

$$\phi(x) = \max(0, x)$$

Non derivabile in 0, altrove espone una dinamica (e, quindi, una derivata) ancora più semplice: è il gradino di Heaviside. La non derivabilità in 0 è un problema che si tenta di aggirare scegliendo una funzione concettualmente simile...

- ...la funzione softplus,

$$\phi(x) = \ln(1 + e^x)$$

la cui derivata esiste ed è pari alla funzione logistica.

La regressione logistica: la rete neurale minima Dato un primo strato di percettroni, che si occupano meramente dell'acquisizione dei dati, tutti connessi ad strato monopercettrone che emette l'output, se la funzione di attivazione del percettrone di output è l'ormai famosa sigmoide, il compito svolto da questa rete neurale minima nient'altro è che la regressione o classificazione logistica.

K-nearest neighbors: sulle metriche di prossimità Un approccio alternativo all'interpretazione di dati multidimensionali si avvale delle nozioni di metrica di prossimità tra punti immersi in \mathbf{R}^n . Come descritto dal famoso eserciziario Marcellini-Sbordone [14]:

"Sia X un insieme e $d : X \times X \rightarrow [0, +\infty)$ una funzione. Si dice che d è una distanza o metrica su X , se sono verificate le condizioni:

1. $d(x, y) = 0 \iff x = y$
2. $d(x, y) = d(y, x) \forall x, y \in X$
3. $d(x, y) \leq d(x, z) + d(z, y) \forall x, y, z \in X$

Se d è una distanza su X si dice che (X, d) è uno spazio metrico [...]."

Le metriche di prossimità più comuni

- Distanza di Minkowski: famiglia di distanze continue

$$d(x, y) = \left(\sum_{i=1}^n |y_i - x_i|^p \right)^{\frac{1}{p}} \quad (2.12)$$

Da notare che, per valori particolari di p , si ottengono molte delle distanze più famose:

1. $p=2$: la distanza euclidea

$$d = \left(\sum_{i=1}^n |y_i - x_i|^2 \right)^{\frac{1}{2}} = \sqrt{(y - x)^2} \quad (2.13)$$

2. $p=1$: la distanza in modulo assoluto:

$$d = \sum_{i=1}^n |y_i - x_i| \quad (2.14)$$

3. $p \rightarrow \infty$: la distanza di Chebyshev

$$d = \lim_{p \rightarrow \infty} \left(\sum_{i=1}^n |y_i - x_i|^p \right)^{\frac{1}{p}} = \max_i \{|y_i - x_i|\} \quad (2.15)$$

- Distanze discrete

Esistono varie nozioni discrete di distanza.

1. Distanza discreta - per antonomasia

$$d(x, y) = (x == y) \quad (2.16)$$

2. Distanza di Hamming

Su un insieme \mathbf{F}^n di stringhe di lunghezza n su un alfabeto \mathbf{F} , $d(x, y) = \text{cardinalità}(\{ i : x_i \neq y_i \})$

Random forest: il modello esplicabile Foresta casuale, dall'inglese, di alberi. L'albero decisionale, in particolare, ovvero una rete di selettori che categorizzano in maniera via via più raffinata i dati di allenamento in ingresso, fino a degli insiemi con una cardinalità minima preimpostata chiamati "foglie". Dopodiché, tocca ai dati di test, per cui ogni albero compirà un errore minimo. La tecnica della foresta casuale è un modello di apprendimento ensemble, poiché inizializza molteplici alberi, tra i quali è possibile riconoscere l'albero migliore (cioè, quello che performa meglio, col minimo errore). La tecnica è di interesse nell'ambito delle intelligenze artificiali spiegabili, ovvero di cui è facile comprendere il processo matematico che mappa gli input negli output. Infatti, basterà ripercorrere il percorso decisionale (la sequenza ottimale di decisioni binarie) per capire la categorizzazione compiuta dal modello. Come con tutti gli alberi, alla struttura può essere richiesto di arrestarsi a varie profondità.

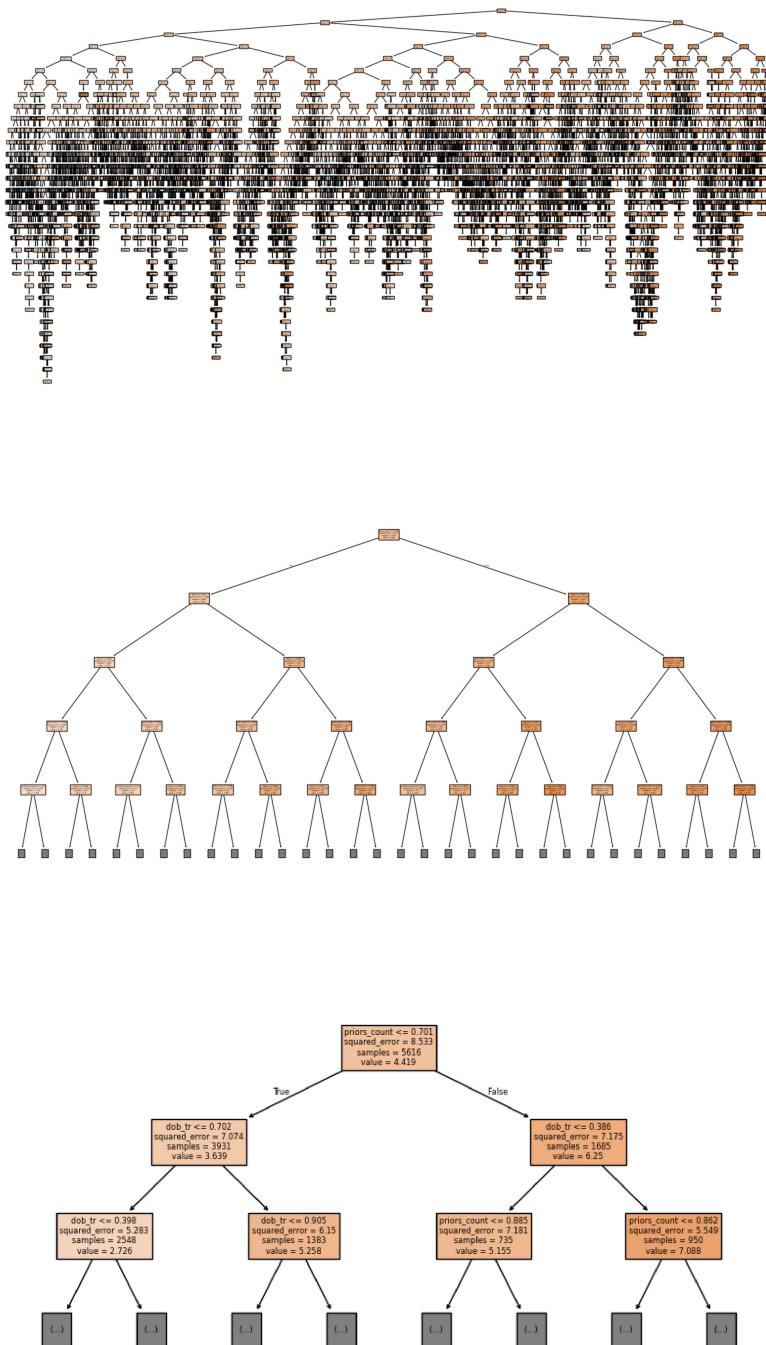


Figura 2.3: L'albero migliore della Foresta, generata con i dati di COMPAS, poi potato a varie profondità (depth=4, depth=2)

Support Vector Machines: il metodo kernelizzato L'idea di base è l'identificazione di un'unico iperpiano nell'iperspazio in cui sono immersi i dati, dal quale è massima la distanza dei punti più prossimi. La ricerca è del più robusto vincolo monolatero (scleronomo, ovviamente) che separi linearmente (in uno spazio eventualmente aumentato di dimensioni extra, se non fossero nativamente linearmente separabili) la collezione di dati che gli arrivano in input nel modo più esatto possibile.

I kernel I tre kernel più famosi sono:

- Il kernel lineare

$$K(x, y) = x^T y \quad (2.17)$$

- Il kernel polinomiale

$$K(x, y) = (x^T y + r)^n \quad (2.18)$$

- Il kernel gaussiano (o radiale, o RBF)

$$K(x, y) = e^{-\frac{|x-y|^2}{2\sigma^2}} \quad (2.19)$$

Ma perché il kernel è così importante per questi modelli? Il kernel genera uno spazio隐式的 in cui non è necessario calcolare esplicitamente nuovi parametri e nuove coordinate. L'approccio, infatti, è quello:

1. di associare l'i-esimo record dei dati di allenamento (x_i, y_i), imparandone un peso scalare w_i
2. e di applicare ai dati di verifica x' una funzione di similarità k (appunto, il kernel)

2.5 Le pipelines e le metriche di performance

Vista la nota best practice del preprocessing, gli sviluppatori della libreria scikit-learn hanno previsto la necessità di associare, in un unico processo, sia il preprocessore dei dati che il modello che si nutre direttamente dell'output del preprocessore.

Poiché ciascun oggetto istanziato da una classe della libreria scikit-learn espone un metodo `.fit()` e un metodo `.predict()`, le pipelines, essendo di classe Pipeline, non fanno eccezione. Delle varie metriche, la metodologia selezionata per questa tesi prevede di comparare varie pipelines, composte ciascuna da un preprocessore seguito da un modello tra quelli esaminati, sulla base del `mean_squared_error` (MSE) summenzionato. Si mostrano anche i risultati del `R2_score` (cioè, il coefficiente di determinazione: $R^2 = 1 - \frac{RSS}{TSS}$, dove RSS sta per la somma residua dei quadrati, mentre TSS è la somma totale dei quadrati). R2 segnala, se è prossimo a 1, il buon comportamento predittivo del regressore, ma la sua espressività è limitata alle regressioni lineari. Le regressioni tra cui si vuole esplorare non lo sono, il che apre a `R2_scores` negativi, che quindi non sono significativi.

Capitolo 3

Conclusioni

3.1 Presentazione dei risultati

Come già spiegato, mse_score coincide con una nozione classica di errore, pertanto di queste 15 tecniche di analisi dei dati risulta interessante cercare quella per cui è stato registrato il minimo errore quadratico medio. Questo primato sembra spettare alla pipeline costituita dal preprocessore che trasforma tramite quantili, fatto seguire dalla Support Vector Machine (svr, perché il suo compito è di regressione). Si presenta comunque la r2_score, sapendo che non si applica per predittori non lineari, proprio per mostrarne il perché.

| Model | Pre-scaler | MSE Score | R2 Score |
|---------------|------------|-----------|----------|
| logistic | standard | 8.7636 | -0.0945 |
| logistic | minmax | 8.9271 | -0.1149 |
| logistic | quantile | 7.4837 | 0.0654 |
| svr | standard | 18.4113 | -1.2994 |
| svr | minmax | 6.5447 | 0.1826 |
| svr | quantile | 4.6640 | 0.4175 |
| knn | standard | 6.4892 | 0.1896 |
| knn | minmax | 7.0871 | 0.1149 |
| knn | quantile | 5.0077 | 0.3746 |
| random_forest | standard | 6.1152 | 0.2363 |
| random_forest | minmax | 7.3495 | 0.0821 |
| random_forest | quantile | 5.1708 | 0.3542 |
| voting | standard | 5.5549 | 0.3062 |
| voting | minmax | 6.9400 | 0.1332 |
| voting | quantile | 4.8342 | 0.3963 |

Tabella 3.1: *Model Performance Comparison*

3.2 Un'ultima indagine sui dati: la PCA

La PCA, Principal Component Analysis, è una tecnica che permette di individuare gli autovettori della matrice di covarianza, o in altre parole quale vettore di pesi nella combinazione lineare dei dati spiega la maggior porzione di varianza possibile. Invece che prendere i dati così come sono, con la PCA è possibile generare un nuovo spazio, a dimensionalità ridotta, i cui dati sono calcolati con la combinando linearmente i dati usando come pesi le componenti dell'autovettore selezionato.

Le features più rilevanti Ciascuna delle tre tabelle in figura 3.1 deriva da una PCA cui sono stati dati in input i dati preprocessati tramite una delle tre tecniche viste in precedenza. La terza tabella mostra risultati di qualità significativamente più bassa, visto che le prime cinque componenti spiegano, in somma, a malapena il 55% della varianza complessiva dei dati in input. Va meglio con le altre due PCA:

- Il quantile transformer spiega il 78%
- Il minmax scaler spiega il 95% (praticamente, tutto)

Ma il risultato più rilevante, è sull'importanza delle features. Sempre restando sulle prime due tabelle, si evidenzia che la componente 1, che spiega rispettivamente il 30 e il 39 % della varianza, è importante per oltre il 70% essere afro-americani. E, per oltre il 60%, non essere caucasici. La recidività è importante solo ad un occhio più attento che si spinga alla seconda componente principale, che però è seconda proprio perché spiega una percentuale inferiore della varianza dei dati (il metodo restituisce le componenti in ordine decrescente). Il metodo rileva correttamente anche un altro fatto, molto meno grave: una correlazione tra la feature "dob_tr" e la feature "age", come prevedibile dalla semantica di queste features.

3.3 Sull'evoluzione della nozione di deviante

Anche prima dell'avvento delle intelligenze artificiali, la sociologia della devianza si interroga da secoli su come capire chi compie devianza e perché. [15] Non tutta la devianza è reato, ma tutto il reato è devianza.¹ Nella storia del pensiero sociologico si annoverano 6 teorie principali:

- Teoria biologica dell'atavismo criminale (Cesare Lombroso, tardo Ottocento)

Uno dei primi studiosi che ha dato una scientificità al desiderio di predire il criminale è stato il medico e psichiatra Cesare Lombroso (1835-1909). Sotto la spinta dell'allora recente teoria darwiniana dell'evoluzione, Lombroso ha ritenuto che il criminale ne fosse rimasto indietro, e ha cercato le prove analizzando le caratteristiche fisiche e biologiche dei criminali, concentrandosi soprattutto sulla forma cranica. Ad ogni modo, la teoria è stata ripetutamente oggetto di critiche, fino a decadere. Esiste il museo di Antropologia criminale, a Torino, dedicato a lui e alla sua teoria. Gli allestitori del museo hanno dichiarato che alla base del museo esiste:

una funzione educativa intesa a mostrare come la costruzione della conoscenza scientifica sia un processo che avanza grazie alla dimostrazione non tanto di verità, quanto della "falsificabilità" di dati e teorie che non resistono a una critica.

Il nuovo allestimento vuole fornire al visitatore gli strumenti concettuali per comprendere come e perché questo personaggio così controverso formulò la teoria dell'atavismo criminale e quali furono gli errori di metodo scientifico che lo portarono a fondare una scienza poi risultata errata.[16]

- Teoria della tensione (Robert K. Merton, anni '60)

Merton (1910-2003), sociologo, scomponete la cultura in mete culturali (ovvero, le prospettive sociali cui tendere, "gli scopi nella vita") e in mezzi istituzionalizzati (i processi sociali autorizzati dalla società e presenti nella società). Merton reinterpreta l'anomia di Émile Durkheim (1858-1917) come tensione (o distanza) tra la struttura sociale istituzionale e la struttura culturale della società. Dunque, individua 5 modi di adattamento alla società:

1. Conformità : accettare entrambi gli aspetti, e dunque non deviare.
2. Innovazione: accettare le mete, ma rifiutare i mezzi. Per esempio si vuole diventare più ricchi, ma magari lavorando in modi innovativi o cambiando carriera (eventualmente, passando ad una carriera criminale).
3. Ritualismo : rifiutare le mete, ma accettando i mezzi. Per esempio, lavorare, ma senza l'ambizione comune di diventare più ricco, e quindi fare appena il necessario.
4. Rinuncia : abbandonare entrambi gli aspetti, e quindi ritirarsi dalla vita sociale in toto (aprendo quindi a fenomeni di marginalità sociale: clochardage, eremitismo - anche urbano, ecc.).

¹Sempre nel testo di Bagnasco-Barbagli-Cavalli [15], a pp. 112-113 si menzionano delle ricerche sociologiche interessate all'analisi della classe sociale, dell'età e del genere dell'autore di reati.

5. Ribellione : rifiutare, ma protestare per la sostituzione di uno o entrambi i valori della società con nuovi.
- Teoria della subcultura (dai critici di Merton)
I vari critici della teoria della tensione ritenevano che non fosse sufficiente a spiegare il comportamento deviante, poiché lo ascriveva ad una mera intenzionalità individuale. Invece, soprattutto i membri più celebri della Scuola di Chicago (fondata da Robert Park (1864-1944)), Clifford Shaw e Henry McKay, hanno condotto mastodontiche ricerche sociologiche alla ricerca delle subculture dove si apprendono sia le competenze tecniche, sia le razionalizzazioni che favoriscono l'insorgenza di attività criminali. L'esempio scolastico è quello dello scasso: servono delle competenze specifiche per lo scasso, nonché dei valori più solidi di quelli che legano l'individuo alla società perché questo individuo inizi a delinquere; non è una cosa innata (Lombroso) o una "invenzione dell'attore" (Merton) ma, in soldoni, secondo questi studiosi la strada è la scuola dei devianti. Uno dei maggiori criminologi americani del Novecento, Edwin Sutherland (1883-1950), riprende questa teoria, e la espande notando, tra le altre cose, che chi sta deviando non sta veramente deviando, se si analizza l'individuo in relazione alla subcultura in cui è immerso: deviante è il gruppo rispetto alla società generale, e non l'individuo, essendo che i valori della subcultura si frappongono tra l'individuo e la società generale.
 - Teoria del controllo sociale (Travis Hirschi)
La teoria della tensione di Merton summenzionata, si basa sull'assunto che l'individuo sia naturalmente portato a seguire mete e mezzi socialmente determinati prima, e accettati poi. La teoria del controllo sociale muove le sue premesse da visione più pessimistica dell'individuo, moralmente molle/debole/flessibile. Pertanto, la prospettiva viene ribaltata: non ci si chiede più perché l'individuo compie reati, bensì come mai non li compia. La risposta è "perché sono frenate dal farlo" dal controllo sociale, cioè dalla sensibilità individuale:
 - al controllo esterno (cioè alla sorveglianza degli altri individui - siano essi rappresentanti di istituzioni pubbliche oppure membri del "gruppo dei pari", cioè del gruppo di amici - per la prevenzione dei comportamenti devianti)
 - al controllo interno diretto (cioè ai sentimenti di imbarazzo, colpa e vergogna interni al trasgressore)
 - al controllo interno indiretto (ovvero all'attaccamento psicologico ed emotivo agli altri, nonché al desiderio di non perdere la loro stima e il loro affetto)
 - Teoria della scelta razionale (Hiroshi Tsutomi)
Le teorie precedenti non prendono in nessuna considerazione l'intenzionalità del reo di violare le norme. Il criminologo giapponese Hiroshi Tsutomi rileva che le persone commettono reati non perché sono patologiche o malvagie, ma perché sono normali. Le motivazioni sono sempre quelle: la ricerca di potere, prestigio, profitto, piacere. Altri sostenitori della teoria sono Cesare Beccaria (tra le tante altre cose, l'autore del

trattato "Dei delitti e delle pene" contro la pena di morte e la tortura) e Jeremy Bentham (uno dei più celebri filosofi utilitaristi). Infatti la teoria ha un chiaro sapore economico (sembra ricalcare abbastanza fedelmente le idee del "voto del portafogli" e del libero convincimento del consumatore in un mercato), ed è ripresa non solo da sociologi anche contemporanei, ma anche da economisti. Viene, appunto, individuata una classificazione dei costi del delinquere:

- Costi esterni pubblici (le sanzioni legali comminate dallo Stato e le sanzioni sociali sulla reputazione)
 - Costi esterni privati (o "costi di attaccamento": le sanzioni informali dagli "altri significativi")
 - Costi interni (emergenti dalla coscienza, o dalle "norme interiorizzate", principalmente sotto forma di senso di colpa e vergogna)
- Teoria dell'etichettamento (dall'interazionismo simbolico)
Premesse:

- William Thomas: "Se gli uomini definiscono reali certe situazioni, esse saranno reali nelle sue conseguenze." (Teorema di Thomas; esempio: banca che viene realmente messa in crisi dalle malelingue sul presunto stato di crisi in cui verserebbe, anche fossero state false dicerie.)
- George Herbert Mead: "I significati e le identità (personal e sociali) emergono tramite le interazioni sociali"
- Charles Cooley (il teorico del "sé specchio"): "Il comportamento degli individui tiene molto conto di ciò che gli altri pensano, della nostra percezione della percezione che gli altri hanno di lui." - l'interazione è centrale e mentalistica.

La teoria

- La definizione basale
Devianza : etichettamento, che resiste nel tempo tramite l'interazione.
E' una definizione molto poco centrata sul deviante, che rovescia la prospettiva sul processo con cui identifichiamo ed etichettiamo il deviante, che quindi è un costrutto sociale.
- Howard Becker, "Outsiders. Saggi di sociologia della devianza", 1963
I gruppi sociali creano la devianza, istituendo norme la cui infrazione costituisce la devianza stessa, applicando quelle norme a certe persone, e a quelle persone l'etichetta di deviante. La devianza non è una qualità dell'atto commesso, piuttosto una conseguenza dell'applicazione, da parte di altri, di norme e sanzioni verso un "colpevole". Il comportamento deviante è il comportamento così etichettato.

– Edwin Lemert

La devianza è comune, e spesso senza conseguenze. Due tipi:

- * Devianza primaria (poligenetica. Anche se socialmente sgradita, la devianza primaria presenta implicazioni marginali per lo status e per la struttura psichica della persona interessata)
- * Devianza secondaria (la scoperta del reato, tramite la stigmatizzazione, porta alla creazione sociale e psicologica del deviante. L'individuo inizia a percepire come deviante, fino a farla diventare la caratteristica principale - cioè, fino a sentirsi più deviante che individuo; potrebbe quindi iniziare una "carriera criminale".)

Bibliografia

- [1] “Se l’amicus curiae è un algoritmo: il chiacchierato caso Loomis alla Corte Suprema del Wisconsin.” (), indirizzo: <https://www.giurisprudenzapenale.com/2019/04/24/lamicus-curiae-un-algoritmo-chiacchierato-caso-loomis-allla-corte-suprema-del-wisconsin/>.
- [2] G. Sartor, F. Lagioia et al., “Il sistema COMPAS: algoritmi, previsioni, iniquità,” in *XXVI Lezioni di Diritto dell’Intelligenza artificiale*, Giappichelli, 2021, pp. 226–244.
- [3] G. Sartor, *Intelligenza artificiale e diritto. Un’introduzione*. Giuffrè editore, 1996, ISBN: 9788814053863.
- [4] “Giustizia predittiva: quale futuro?” (), indirizzo: <https://www.altalex.com/documents/news/2023/03/15/giustizia-predittiva-quale-futuro>.
- [5] “Giustizia predittiva: dubbi tra innovazione e regresso.” (), indirizzo: <https://www.altalex.com/documents/2024/01/04/giustizia-predittiva-dubbi-innovazione-regresso>.
- [6] L. Viola, *Interpretazione della legge con modelli matematici. Processo, a.d.r., giustizia predittiva. Volume I*, 2^a ed. Centro Studi Diritto Avanzato Edizioni, 2018, ISBN: 9788828388708.
- [7] H. F. Julia Dressel, “The accuracy, fairness, and limits of predicting recidivism.,” 1996.
- [8] “Machine bias: risk assessments in criminal sentencing.” (), indirizzo: <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>.
- [9] “Artificial Intelligence: current Italian regulations and future perspectives.” (), indirizzo: <https://www.brocardi.it/tesi-di-laurea/artificial-intelligence-current-italian-regulations-future/819.html>.
- [10] “COMPAS analysis.” (), indirizzo: <https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm>.
- [11] “COMPAS analysis - GitHub resources.” (), indirizzo: <https://github.com/propublica/compas-analysis/tree/master>.
- [12] P. A. et al., *Basi di dati*, 5^a ed. McGraw-Hill, 2018, ISBN: 9788838694455.
- [13] “Bias Variance Tradeoff.” (), indirizzo: <https://www.cs.cornell.edu/courses/cs4780/2018fa/lectures/lecturenote12.html>.
- [14] C. S. Paolo Marcellini, *Esercizi di Matematica, Volume II, Tomo I*. Liguori editore, 2009, ISBN: 9788820746490.
- [15] A. C. Arnaldo Bagnasco Marzio Barbagli, *Elementi di sociologia*, 2^a ed. Il Mulino Editore, 2013, ISBN: 9788815241825.
- [16] “Il museo di antropologia criminale.” (), indirizzo: <https://www.museolombroso.unito.it/museo/intro/>.

Appendici

Appendice A

test_sqlite.py

```
import sqlite3
from datetime import date

#### Date handler (conversion to integers)

pot_dates = lambda giorno: (giorno - date(1936, 6, 24)).days # Ciao, nonna.
date_converter = lambda item: \
    None if item == None or item == '' \
    else pot_dates(date(*[
        int(i) for i in (
            item[0:4],
            item[5:7],
            item[8:10]
        )
    ]))

#### Utilities and Loading data

file = "data\\compas-analysis\\compas.db"
con = sqlite3.connect(file)
cur = con.cursor()
gen = lambda col: cur.execute(
    f"SELECT {col} FROM people WHERE is_recid != -1"
)
gen_list = lambda gen: [val for tup in gen for val in tup]
date_int = lambda lst: list(map(date_converter, lst))
table_gen = lambda data: {col : gen_list(gen(col)) for col in data}
lindt = [*cur.execute("PRAGMA table_info('people')")]
columns = [lindt[_][1] for _ in range(len(lindt))]
data = {col : gen_list(gen(col)) for col in columns}

none_number = lambda caller, data, columns: \
print(*[
```

```
f"{{ caller }}" + \
f"{{ i }}:" + \
f"Nones: {{ data[i].count(None)}}" + \
f"Uniques: {{ len(set(data[i]))}}" + \
f"Total: {{ len(data[i])}}"
for i in columns
], "", sep="\n")
people = ('id', 'first', 'last')
people_data = table_gen(people)
```

Proiezione algebrica dei dati

```
short_columns = (
    'id',
    'sex',
    'race',
    'dob',
    'age',
    'juv_fel_count',
    'juv_misd_count',
    'juv_other_count',
    'priors_count',
    'is_recid',
    'decile_score',
)
short_data = table_gen(short_columns)
# projected data structure
# none_number('projected_data:', short_data, short_columns)
assert data['is_recid'].count(-1) == 0

race_uniques = [*set(short_data['race'])]
race_list = [f"{{ i.lower()}}" for i in race_uniques]
transformed = ('is_male', 'dob_tr', *race_list)
# data['id'] = [data['id'][i] - 1 for i in range(len(data['id']))]
short_data['id'] = [item - 1 for item in short_data['id']]
assert (short_data['id'][i] == i for i in range(len(short_data['id'])))
short_data['is_male'] = [int(item == 'Male') for item in short_data['sex']]
short_data['dob_tr'] = [date_converter(item) for item in short_data['dob']]
for i in range(len(race_uniques)):
    short_data[race_list[i]] = \
        [int(item == race_uniques[i]) for item in short_data['race']]
# revised columns
# none_number('transformed_data:', short_data, transformed)
kys = [*short_data.keys()]
#print(kys)
kys.remove('sex')
kys.remove('race')
```

```
kys.remove('dob')
kys.append(kys.pop(6))
kys.append(kys.pop(6))
#print(*kys, sep = '|n')
ultimate_data_dict = {i : short_data[i] for i in kys}
for key, values in ultimate_data_dict.items():
    for value in values:
        assert type(value) == int
ultimate_data_list = [*ultimate_data_dict.values()]
# print(ultimate_data_list)
```

Appendice B

test_sklearn.py

```
from test_sqlite import *
from prettytable import PrettyTable
import matplotlib.pyplot as plt

from sklearn.preprocessing import *
from sklearn.linear_model import LogisticRegression
from sklearn.neighbors import KNeighborsRegressor
from sklearn.ensemble import RandomForestRegressor, VotingRegressor
from sklearn.svm import SVR
from sklearn.model_selection import train_test_split
from sklearn.pipeline import make_pipeline
from sklearn.decomposition import PCA
from sklearn.metrics import r2_score, mean_squared_error, make_scorer
from sklearn.tree import plot_tree

Z = ultimate_data_dict
Z.pop('id')
y = Z.pop('decile_score')
X = list(map(list, zip(*Z.values()))) # Transpose the list

X_train, X_test, y_train, y_test= \
    train_test_split(X,y, test_size=0.2, random_state=40)

scalers={
    'standard': StandardScaler(),
    'minmax': MinMaxScaler(),
    'quantile': QuantileTransformer()
}

models = {
    'logistic': LogisticRegression(max_iter=1000),
    'svr': SVR(),
    'knn': KNeighborsRegressor(),
```

```

'random_forest': RandomForestRegressor(),
'veting': VotingRegressor( estimators=[
    ('knn', KNeighborsRegressor()),
    ('random_forest', RandomForestRegressor()),
])
}

pipelines = {
    (s, m): make_pipeline( scalers[s], models[m] ).fit(X_train, y_train)
        for m in models
        for s in scalers
}

scorers = {
    'mse': make_scorer(mean_squared_error),
    'r2': make_scorer(r2_score)
}

t = PrettyTable(["Model", "Pre-scaler", "mse_score", "r2_score"])
t.add_rows([
    [
        key[1],
        key[0],
        f'{scorers["mse"]}(pipeline, X_test, y_test):.4f',
        f'{scorers["r2"]}(pipeline, X_test, y_test):.4f'
    ] for key, pipeline in pipelines.items()
])
print(t)

def PCA_shower(n_components, preprocessor):
    print(f'{preprocessor.__name__}'.capitalize())
    pca = PCA(n_components=n_components)
    pca.fit(preprocessor(X_train))
    t = PrettyTable(
        ["Feature"] + \
        [f"Component_{i+1}" for i in range(len(pca.components_))]
    )
    t.add_row([
        "Explained_variance_ratio"] + \
        [
            f'{pca.explained_variance_ratio_[i]:.4f}' \
            for i in range(len(pca.explained_variance_ratio_))
        ], divider=True)
    t.add_rows([[feature] + [f'{pca.components_[j][i]:.4f}' \
        for j in range(len(pca.components_))]] \
        for i, feature in enumerate(ultimate_data_dict.keys())))
    return t

```

```
print(PCA_shower(5, quantile_transform))
print(PCA_shower(5, minmax_scale))
print(PCA_shower(5, scale))

get_rfr = lambda pipeline: pipeline.named_steps['randomforestregressor']
random_forest=get_rfr(pipelines[('standard', 'random_forest')])
tree_scores=[
    (i, mean_squared_error(y_test, tree.predict(X_test)))
    for i, tree in enumerate(random_forest.estimators_)
]
tree_scores_sorted=sorted(tree_scores, key=lambda x: x[1])
best_tree_index, best_tree_mse=tree_scores_sorted[0]
best_tree=random_forest.estimators_[best_tree_index]
print(f"best_tree_index:{best_tree_index}")
print(f"MSE_best_tree:{best_tree_mse:.4f}")

plt.figure(figsize=(10,5))
plot_tree(best_tree, max_depth=2, feature_names=[*Z.keys()], filled=True)
plt.show()
```

ringraziamenti

Vorrei dedicare quest'ultima pagina alle persone che mi hanno sostenuto in questo cammino multilineare, arricchendolo.

Innanzitutto, ringrazio il relatore, Tiziano Politi, per il supporto offertomi nella stesura dell'elaborato, nonostante le vicissitudini tecniche e organizzative siano state copiose.

Colgo anche l'occasione per ringraziare il dott. Piero Marra, Assistant Professor di Filosofia del Diritto presso la facoltà di Giurisprudenza dell'Università degli Studi di Bari, per avermi fatto conoscere il contesto da cui provengono i dati, e le principali implicazioni etiche e giuridiche correlate.

Altro ringraziamento speciale va al prof. Paolo Iagulli, docente di Storia del pensiero sociologico e Sociologia generale del corso di laurea interateneo in Scienze del servizio sociale e Sociologia (L-39 - L-40) dell'Università degli Studi di Bari, per le nozioni di Sociologia della devianza.

Si ringraziano anche il prof. Giuseppe Loseto per Basi di Dati, e il prof. Tommaso di Noia per il primo approccio allo studio delle intelligenze artificiali.

Importante riconoscere il forte impegno personale, emotivo ed economico, ciascuno secondo le proprie inclinazioni e possibilità, che i miei genitori hanno profuso nella creazione di tutti i risultati che hanno portato a questa tesi e a questa discussione di tesi.

Imprescindibile anche il supporto del nonno, della zia e della nonna, di nuovo ciascuno secondo le proprie possibilità.

E ringrazio naturalmente i tanti amici che sono stati vicino in questa esperienza di laurea triennale.